

# A Novel Framework for Future Natural Language Processing From a Database Perspective

Limin Zhang  
liminzhang08@qq.com

October 31, 2023

## Abstract

Most research and applications on natural language still concentrate on its superficial features and structures. However, natural language is essentially a way of encoding information and knowledge. Thus, the focus should be on what is encoded and how it is encoded. In line with this, we suggest a database-based approach for natural language processing that emulates the encoding of information and knowledge to build models. Based on these models, 1) generating sentences becomes akin to reading data from the models (or databases) and encoding it following some rules; 2) understanding sentences involves decoding rules and a series of boolean operations on the databases; 3) learning can be accomplished by writing on the databases. Our method closely mirrors how the human brain processes information, offering excellent interpretability and expandability.

## 1 Introduction

Statistical learning approaches have made impressive progress in natural language processing (NLP) tasks. Large language models (LLMs) are one of the leaders, for example, Transformer [15], BERT [5], GPT-3 [3], etc. However, as black-box models, the effectiveness of LLMs can only be tested in terms of the output results, not by observing their internal working mechanism. This raises concerns and doubts: Can LLMs really learn to understand natural language [2]? This suspicion is reinforced by the hallucination [11, 7], the unfaithful to fact, and nonsensical text generated by LLMs. From our perspective, the correlations and sequence features that LLMs learn in corpora may, to some extent, indicate there are relationships between tokens in different dimensions. However, they are not capable of further discriminating and defining the relationships. Thus, LLMs are just imitating.

In contrast to the approaches that rely on the computing power of machines (i.e., LLMs), linguists have put forward many constructive hypotheses and assertions by taking a closer look at natural language. For example, Chomsky [4] introduced a series of new ideas of mentalistic perspectives and methods of mathematical logic into linguistic research. Talmy [14] regards natural language as a cognitive system, and he recommends focusing on the relationship between deep structure and surface representations. In cognitive grammar [9, 10], Langacker provided a research methodology, which can be summarized in six phases: observation, data collection, theoretical modeling, model verification, and formalization [12].

Inspired by the above thinking and methodology, we extend our focus from natural language alone to the connection between language and the real world, and the relationship between language and human beings. Accordingly, we propose a database-based NLP method that aims to study and reveal how information and knowledge are organized and stored in human brains, then simulate the findings (structures) to construct models (databases), and finally provide the solutions of NLG and NLU (including the learning part) tasks based on these models. Our contributions are threefold.

- Our method changes the study object from natural language itself to the information and knowledge encoded and represented by natural language, giving the model we construct excellent interpretability and expandability.
- We propose a brand new NLP approach that is different from rule-based and statistical model-based (i.e., language model) approaches, which mirrors how the human brain processes natural language.
- Based on the different and deeper understanding of natural language, we create a new framework for studying NLP problems, as well as bringing new thoughts for artificial intelligence (AI) research.

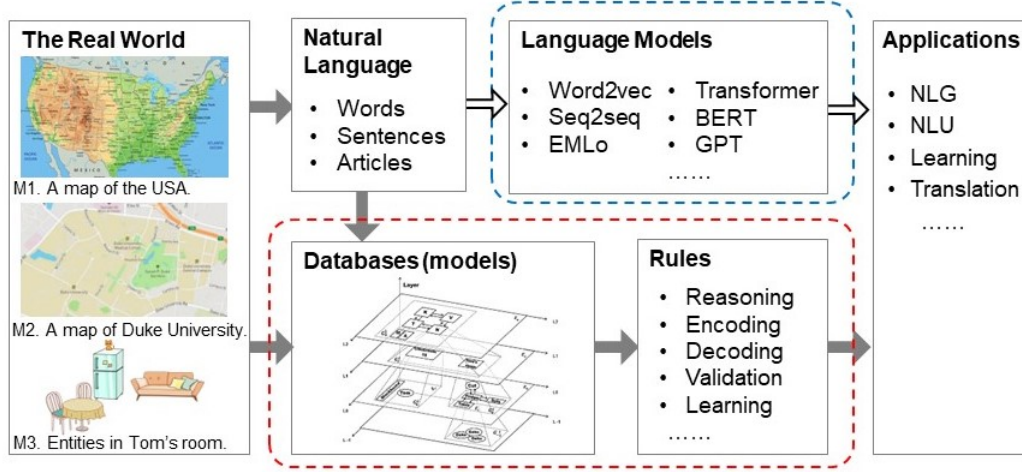


Figure 1: Differences in modeling objects. Language model-based methods take natural language as the object to construct models. Database-based methods take both the information and knowledge in the real world and how they are encoded in natural language as the objects to construct the model.

The United States <i>is south of</i> Canada.	Duke University <i>is in</i> North Carolina.
The cat <i>is in</i> Tom’s room.	The table <i>is in front of</i> the fridge.
The cat <i>is on top of</i> the fridge.	The sofa <i>is next to</i> the fridge.

Table 1: Examples of sentences that describe the spatial position of target entities in the real world

## 2 Overview

As depicted within the red box in Figure 1, our method consists of two parts: 1) Databases and 2) Rules. The databases part is devoted to finding out how information and knowledge are structured and retained in the human brain, accordingly to simulate these structures to build data structures that can be used to generate databases. The rules part simulates how the information and knowledge are processed and utilized in the human brain to accomplish related human-like intelligent activities.

As we learned in neuroscience, humans continuously receive sensory input from the external world, conveyed through various neural pathways, including eyes, ears, noses, and so on[1]. This constant influx of information creates disparities in knowledge and understanding among individuals. To bridge these gaps, humans employ language as a primary tool, which involves the transformation of the information and knowledge stored in the human brain into a communicable form for sharing.

Before language can be generated, there are several essential steps to be taken, including deciding what information and knowledge need to be conveyed, engaging in reasoning (which involves preprocessing meta-knowledge or information), and encoding the knowledge and information. It is crucial to note that different types of information and knowledge encoded in language must be modeled and processed according to their different nature and characteristics. In this paper, we only take the spatial position information as an example to demonstrate the database-based method.

## 3 Method

### 3.1 Preliminaries

People convey spatial position information or knowledge of real-world entities by encoding them into sentences, as shown in Table 1. Looking at these sentences, we can see that they have the same structure: (Entity 1) + (...) + (Spatial relation) + (Entity 2). In this structure, “Entity 1” refers to the **target entity** whose spatial position is being described in the sentence, “Entity 2” serves as a **helper entity** that helps to pinpointing the location of the target entity, and “Spatial relation” describes the spatial relation between the target entity and the helper entity.

Table 2 shows three types of spatial relations commonly used in languages: 1) spatial range relations, 2) spatial directional relations, and 3) spatial distance relations. The spatial directional relations can be further divided into 2.1) absolute directional relations and 2.2) relative directional relations according to the different reference systems.

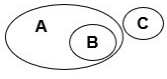
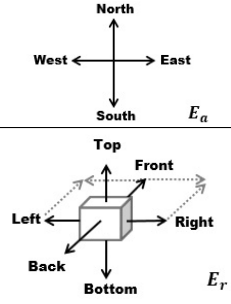
Spatial relations	Lexical representations	Reference system
1.Range relations	<b>Inside:</b> in, at... <b>Outside:</b> outside of...	
2. Directional relations	<b>East:</b> east of... <b>West:</b> west of... <b>North:</b> the north side of ... <b>South:</b> the south side of... <b>Top:</b> on, above, over, on top of... <b>Bottom:</b> below, under, beneath... <b>Left:</b> left of... <b>Right:</b> the right side of... <b>Front:</b> before, in front of... <b>Back:</b> behind, back of...	
3.Distance relations	by, near, next to, beside...	

Table 2: Classification of the relative spatial relations between entities, and lexical representations of the relative spatial relations.

The above findings in language reveal how the knowledge (i.e., the spatial position of real-world entities) is organized and stored in the human brain. We can also see that people are used to using entities with a relatively stable spatial position (immovable entities) as the helper entities. These immovable entities and the spatial relations among them form a stable system, which will serve as the foundation for our model.

## 3.2 TGHM

We construct a TGHM (**T**ree **G**raph **H**ybrid **M**odel) to describe and store the spatial position of real-world entities. In a TGHM, the real-world entities are abstracted as nodes; the spatial relations between these nodes are abstracted as directed edges  $E$ . The TGHM itself can be understood as a data structure composed of two fundamental components: Tree and Graph.

### 3.2.1 Tree Model

We use a tree model to describe the **spatial range relations** ( $E_s$ ) between entities.  $E_s$  is consist of two opposite directions, i.e.,  $E_s = \{\overrightarrow{inside}, \overleftarrow{outside}\}$ . For example, we use the tree in Figure 2 to describe the spatial range relations between entities “North Carolina”, “Duke University”, “Tom’s room”, “Table”, “Cat”, etc. The tree in Figure 2 can also be written in tabular form as shown in Table 3. In a tree, child nodes with a same parent node should be spatially independent of each other, which means, there is no spatial range inclusion relation between them, if not, the child node must be moved up or down until all the child nodes are spatially independent of each other.

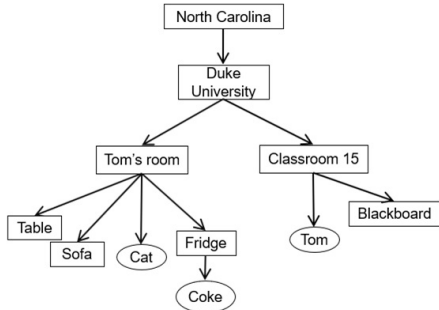


Figure 2: The tree that describes the spatial range relations between entities “North Carolina”, “Duke University”, “Table”, “Cat”, etc.

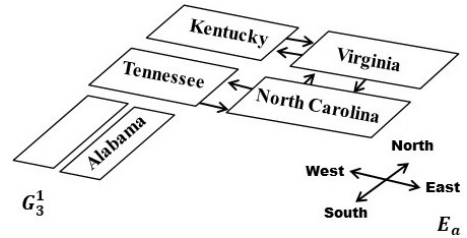


Figure 3: The graph that describe the absolute spatial directional relations between some entities in M1 in Figure 1.

$E_s \backslash V$	North Carolina	Duke University	Tom's room	Classroom 15	Fridge
$\xrightarrow{inside}$	Duke University	Tom's room, Classroom 15	Table, Sofa, Cat, Fridge	Blackboard, Tom	Coke
$\xleftarrow{outside}$	$\emptyset$	North Carolina	Duke University	Duke University	Tom's room

Table 3: The tabular form of the tree in Figure 2

### 3.2.2 Graph Model

We use graph models to describe the spatial directional relations between entities. The **spatial directional relations** can be future divided into **1) absolute directional relations** ( $E_a$ ), which consists of four fixed directions, i.e.,  $E_a = \{\overrightarrow{east}, \overrightarrow{west}, \overrightarrow{north}, \overrightarrow{south}\}$ , and **2) relative directional relations** ( $E_r$ ), which consists of six fixed directions, i.e.,  $E_r = \{\overrightarrow{left}, \overrightarrow{right}, \overrightarrow{front}, \overrightarrow{back}, \overrightarrow{top}, \overrightarrow{bottom}\}$ . Now, we can use the graph in Figure 3 to describe the absolute directional relations between some entities in M1 in Figure 1, and use the graph in Figure 4 to describe the relative directional relations between the entities in M3 in Figure 1. These two graphs can also be written in tabular forms.

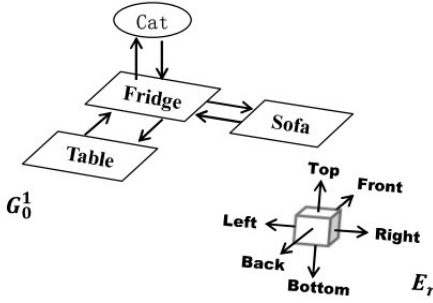


Figure 4: The graph describes the relative spatial directional relations between the entities in M3 in Figure 1

$E_r \backslash V$	Table	Fridge	Sofa
$\xrightarrow{left}$	$\emptyset$	$\emptyset$	Fridge
$\xrightarrow{right}$	$\emptyset$	Sofa	$\emptyset$
$\xrightarrow{front}$	Fridge	$\emptyset$	$\emptyset$
$\xrightarrow{back}$	$\emptyset$	Table	$\emptyset$
$\xrightarrow{top}$	$\emptyset$	Cat	$\emptyset$
$\xrightarrow{bottom}$	$\emptyset$	$\emptyset$	$\emptyset$

Table 4: The tabular form of the graph in Figure 4

### 3.2.3 TGHM

Finally, taking the nodes common to the tree and the graphs in Figures 2, 3 and 4 as connection points, we can integrate the tree and the graphs into the tree-graph hybrid model (TGHM) as shown in Figure 5. TGHM describes the spatial range relations between entities on the vertical structure (i.e., the inter-layer structure); and the spatial directional relations between entities on the horizontal structure (i.e., the intra-layer structure). Each layer of a TGHM can accommodates multiple subgraphs. Usually, the  $E_a$  (absolute directional relations) is used as the reference frame of the whole layer, and the  $E_r$  (relative directional relations) is used as the reference frame in each subgraph. As shown in Figure 5, the subgraph  $G_0^2$  and  $G_0^1$  take  $E_r$  as their reference frame, and the layer L0 is using  $E_a$  as its reference frame. In a TGHM, the immovable nodes and the edges between them form a stable frame, which is a new reference frame in addition to the widely used numerical positioning system (e.g., GPS). TGHM also provides a bridge for information (knowledge) exchange between language and the numerical reference frame, as shown in Figure 6.

TGHM offers remarkable flexibility in its structural expansion. Vertically, it can be continuously extended to add nodes, and horizontally, it can be subdivided to accommodate additional nodes. This adaptability empowers TGHM to satisfy people's need to describe and store the spatial position of numerous entities in the real world. When an entity's spatial position changes, the TGHM's corresponding data is easily adjusted to reflect these modifications. In addition, we can also build datasets to store the spatial position of the movable entities to record their footprint. Consequently, TGHM enables us to simulate how humans organize and store the spatial position of real-world entities in the brain, which means we can create memories for machines. Coupled with

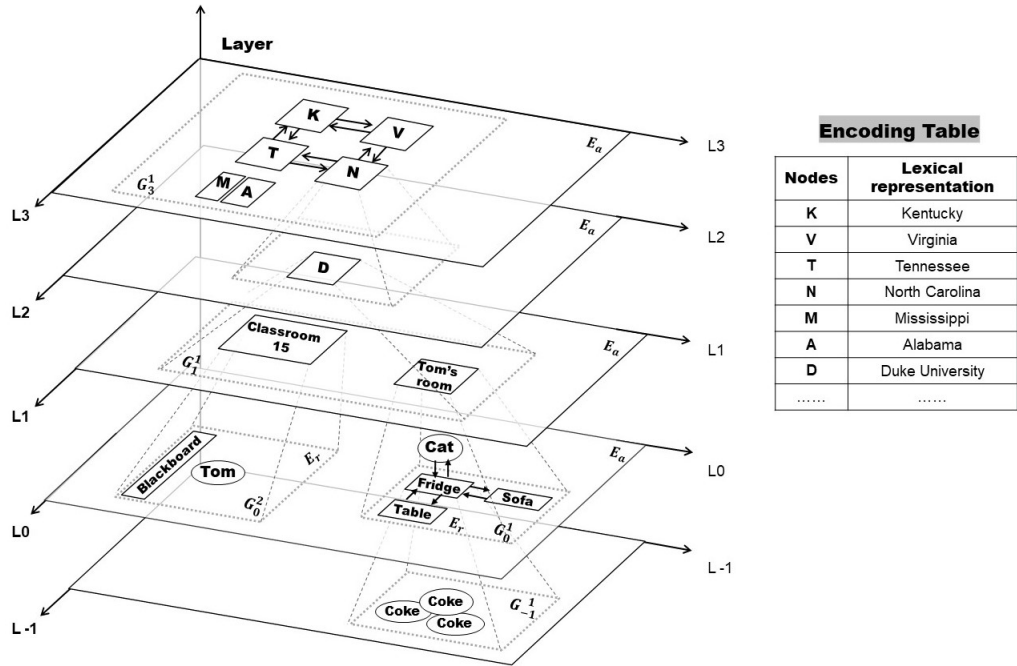


Figure 5: A perspective example of the TGHM, which describes the spatial range relations between entities in the vertical structure (inter-layer structure) and the spatial directional relations between entities in the horizontal structure (intra-layer structure). The encoding table stores the lexical representations in different languages corresponding to each node in the TGHM. Here, we only list English as an example.

pertinent data processing rules, this capacity equips machines to manage and apply the data contained within the TGHM, representing a practical application of human-like intelligence.

## 4 Data Processing Rules

TGHM is consistent with the characteristics of the structural-mechanistic kind of model [13], which follows an underlying mechanistic understanding of reality. Therefore, the TGHM can be used for many purposes. In this paper, we only present how the data in the TGHM has been processed and utilized in NLG and NLU (including the learning part) tasks.

### 4.1 Natural Language Generation

Since language is a tool used to convey information and knowledge, the NLG task can be further broken down into **subtask 1**- determining the information and knowledge that needs to be conveyed, and **subtask 2**- encoding that information and knowledge into sentences.

#### 4.1.1 Data Reading

In this paper, the knowledge to be conveyed is the spatial position of the target entity. We adhere to linguistic conventions employing a helper entity in conjunction with spatial relationships between this helper and the target entity to describe the spatial position of the target entity. For example, if we intend to describe the spatial position of the entity “Cat”, the initial step involves identifying the corresponding node (target node) of the entity “Cat” in TGHM in Figure 5. Subsequently, to identify the helper nodes that have a spatial relation with the target node, such as the nodes “Tom’s room”, “Fridge”, “Table”, “Duke University”, “Tennessee”, and so on. After this process, we can gain 5 distinct data chains as shown in Figure 7, which are composed of the target node, the helper node, and the spatial relations between them.

Each of these 5 data chains can describe the spatial position of the entity “Cat”; their precision varies. If we sort these 5 data chains by precision, we can get the following result:  $L2 > L3 > L1 > L4 > L5$ . However, precision is not the only goal we are pursuing when generating a sentence. Suppose our goal is to communicate the spatial position of the entity “Cat” to a specific individual. In that case, we must also consider the person’s existing knowledge about the spatial positions of the 5 potential helper nodes and their requirements for descriptive precision to filter out the appropriate one accordingly. We will skip this part and go directly to the



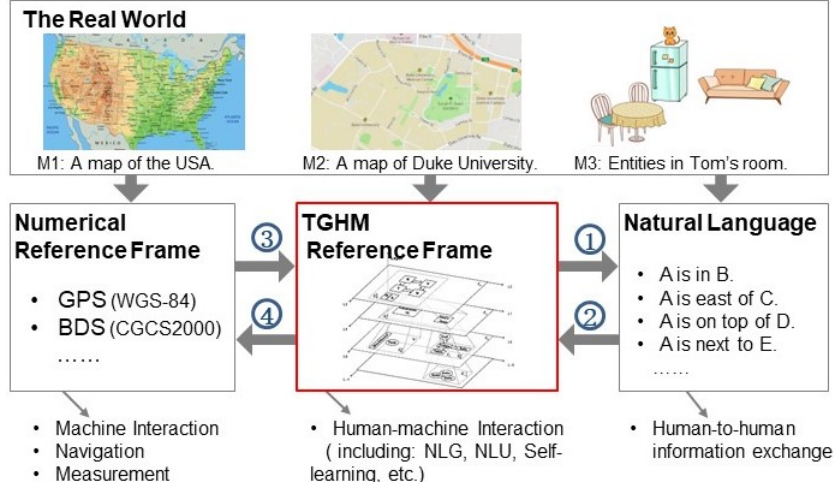


Figure 6: Three ways of describing (or encoding) the spatial position of real-world entities. The information exchange routes between different systems: ① sentences generation. ② sentences understanding. ③ search for neighboring entities. ④ get the numerical position of the target entities.

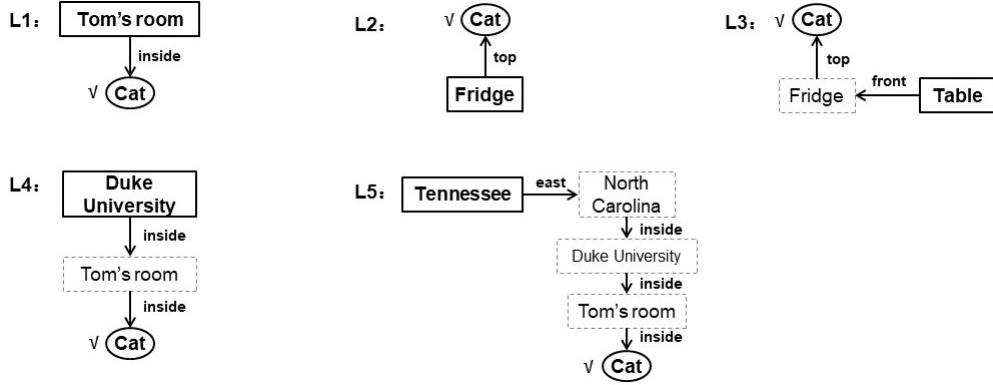


Figure 7: If we take the entity “Cat” as the target, then we can find the above 5 data chains in the TGHM to help locate the entity “Cat”. The nodes marked with "✓" are the target nodes.

sentence encoding phase. Filtering out the appropriate helper node and reading out the data chain is the goal of **subtask 1**.

#### 4.1.2 Encoding Rules

While the specific rules for encoding a data chain into a sentence may exhibit slight variations across different languages, certain essential components remain consistent. These include: 1) the target and the helper nodes in the data chain, 2) the spatial relation between the target node and helper node, and 3) existence judgment of the spatial relation.

**Existence judgment of the spatial relation** A specific spatial relation between two nodes can either be present or absent. In language, we use "be" and "be not" to denote these two conditions. For instance, the words "is" and "is not" in Row 4 of Table 5 describe whether the spatial relation in Row 3 exist or not.

**Reasoning of the spatial relations** In the TGHM, the spatial relation between any two nodes can be calculated. We have summarized some reasoning rules according to how humans process them; see examples below:

- *Elimination operation:* e.g.,  $\xrightarrow{inside} + \xleftarrow{outside} = \emptyset$ ,  $\xrightarrow{left} + \xrightarrow{right} = \emptyset$ ,  $\xrightarrow{north} + \xrightarrow{south} = \emptyset$ ...
- *Union operation:* e.g.,  $\xrightarrow{inside} + \xrightarrow{inside} = \xrightarrow{inside}$ ,  $\xrightarrow{east} + \xrightarrow{east} + \xrightarrow{north} = \xrightarrow{northeast}$ ...
- *Hybrid operation:* when a data chain contains both spatial range relations and spatial directional relations, the relations in the upstream of the data chain is dominant, e.g.,  $\xrightarrow{inside} + \xrightarrow{top} = \xrightarrow{inside}$ ,  $\xrightarrow{east} + \xrightarrow{inside} = \xrightarrow{east}$ ...

If only one edge exists in a data chain, we can encode it directly, such as the data chains L1 and L2. If there is more than one edge in a data chain, e.g., the data chains L3, L4, and L5, we can perform the reasoning rules

Data Chain		L1	L1*	L2	L2*
Main Parts					
1	Target node	The cat	The cat	The cat	The cat
2	Helper node	Tom's room	Tom's room	the fridge	the fridge
3	Spatial relation (E)	in	<i>on top of</i>	on top of	<i>in</i>
4	Existence judgment of the E:				
	• True	is		is	
	• False		<i>is not</i>		<i>is not</i>

Table 5: Examples of the essential components for encoding a data chain.

Data Chain	Target Node	Existence Judgment of the E	Spatial Relation	Helper Node
L1	The cat	is	in	Tom's room
L2	The cat	is	on top of	the fridge
L3	The cat	is	in front of ( <i>next to</i> )	the table
L4	The cat	is	in	Duke University
L5	The cat	is	on the east side of	Tennessee
L1*	The cat	<i>is not</i>	<i>on top of</i>	Tom's room
L2*	The cat	<i>is not</i>	<i>in</i>	the fridge

Table 6: Examples of sentence encoding for the data chains in Figure 7. All the above sentences are 100% correct, but some of them might be regarded as the correct nonsense and will not be adopted in practice due to their low precision in locating the target entity.

to get the results below.

$$L3: \xrightarrow{front} + \xrightarrow{top} = \xrightarrow{upfront}; \quad L4: \xrightarrow{inside} + \xrightarrow{inside} = \xrightarrow{inside}; \quad L5: \xrightarrow{east} + \xrightarrow{inside} * 3 = \xrightarrow{east}.$$

**Distance relations** In some cases, e.g., 1) the spatial distance between the target entity and the helper entity is very close, or 2) it is not necessary to provide the exact position of the target entity, then we can use the spatial distance relations as an alternative, just like the sentence L3 in Table 6

You may argue that the sentences we generated are too simple. However, at the initial stage of language appearance, it is just some simple words and short sentences. With the development of human beings, more and more information is encoded in language, then sophisticated words and long sentences emerge. Therefore, it is a good start to launch our research with some simple words and sentences.

#### 4.1.3 Encoding Rules for Processing Requests

Sentences encode not only the data chain to be conveyed but also the processing requests for that data chain. Based on these implicit processing requests, we divided sentences into the following three categories: 1) data description sentence (i.e., declarative sentence), 2) data verification sentence (i.e., the yes-no question sentence), 3) data searching sentence (i.e., WH-question sentence).

- **Data description sentences** imply the processing request that listeners are expected to store the information or knowledge in their databases. For example, teachers expect the students to remember what was taught in class, and authors expect the readers to understand and remember the ideas shared in the book, and so on.

- **Data verification sentences** imply the processing request that listeners are expected to help verify whether the *spatial relation* described in the sentence is true, and feedback on the verification result as the response. For example, in the case of the data chain L6-1 in Figure 8, speakers are not sure whether the spatial relation “inside” between the node (Duke University) and the node (North Carolina) exists. They could express the processing request that asks listeners to help verify whether the “inside” edge exists by moving the word “Is” to the beginning and adding a question mark at the end of the sentence, as shown in Table 7.



Figure 8: The data chain L6 and its three different cases.

Data Chain		Target Node			Helper Node	
L6		Duke University	is	in	North Carolina	.
L6-1	Is	Duke University		in	North Carolina	?
L6-2	Which state is	Duke University		in		?
L6-3	Which	University	is	in	North Carolina	?

Table 7: Comparison of sentence structures that encode different information processing requests. (English only)

- In **Data searching sentences**, listeners are expected to search for the missing information or knowledge replaced by WH words in their databases and return the search result as the response. Take data chains L6-2 and L6-3 in Figure 8 as examples; speakers can use the word “which” to replace the missing parts and adjust the structure of the sentences, as shown in rows L6-2 and L6-3 in Table 7, to express their expectation that the listener can help to search for the missing parts and return the search results.

## 4.2 Natural Language Understanding

We can consider the NLU process as the inverse of the NLG process. Based on this, the sentence understanding task consists of two parts: a) understanding of the **processing requests** implicit in a sentence, and b) understanding of the **specific knowledge** conveyed in the sentence, which includes all the essential components listed in Table 5. Whereas we only provide the model (TGHM) to describe the spatial relation between entities, here we only introduce the principles for understanding 3) *the spatial relation* and 4) *the existence judgment of the spatial relation*. The understanding of 1) *the target nodes* and 2) *the helper nodes* requires other databases, which are beyond the scope of this paper and will be introduced in other papers in the future.

### 4.2.1 Decoding Rules

**Extract the processing requests** Specific sentence structures, specific feature words, and specific punctuation express the specific processing requests. These can be used to classify the sentences and extract the processing requests accordingly.

**Sentence chunking** Listeners need to chunk the sentence and extract the requisite parts of the specific knowledge. Considering the difference in the number of words and phrases used to represent each class of the

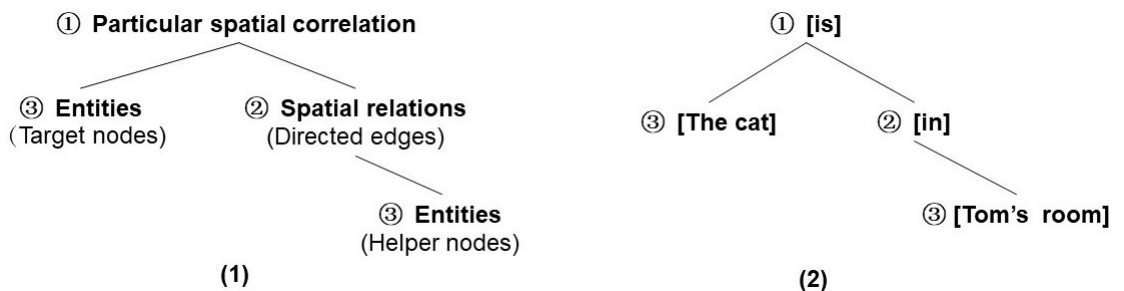


Figure 9: (1) General tree structure of sentences. (2) The sentence tree of sentence L1 in Table 6.



requisite parts, the most efficient way is to chunk the sentences following the order shown in the left part of Figure 9 and gain a sentence tree (see the example shown in the right part in Figure 9).

#### 4.2.2 Understanding of the Specific Knowledge

**Validation** In the process of understanding the specific knowledge, listeners need to verify each part of the sentence tree in their TGHMs according to the flowchart shown in Figure 10. For example, in the case of the sentence tree in Figure 9, listeners should first verify whether the helper entity “Tom’s room” exists in their TGHMs. If the helper entity exists, go ahead; if not, it means that the listeners cannot get the position of the target entity “The cat” through the helper entity “Tom’s room”, so the understanding mission fails. If the helper entity “Tom’s room” exists, the listeners can further verify whether the target entity “The cat” exists at the other end of the “inside” edge. If the target node “The cat” exists, it is known knowledge to the listeners.

**Learning** If the target node “The cat” does not exist, the listeners can create a node at the other end of the “Inside” edge to store the spatial position of “The cat” in their TGHMs. This is a learning process.

**Conflict checking** Furthermore, suppose the specific knowledge described in the sentence conflicts with the knowledge (or data) stored in their TGHMs; a conflict check is required, which may create a new NLG task.

#### 4.2.3 Responding to the Processing Requests

Strictly speaking, responding to the processing requests implies that a sentence serves as sentence-generation task rather than a sentence-understanding task. In this context, we briefly introduce the rules for addressing different processing requests:

- **Data Description Sentence** (i.e., Declarative Sentence):

In response to a data description sentence, such as the one depicted below the dotted line in Figure 10, the reply can be further segmented into validation, learning, conflict checking, and other relevant components, as previously explained.

- **Data Verification Sentence** (i.e., Yes-No Question Sentence):

When dealing with data verification sentences, the primary objective is to provide verification results to the speaker. For example, consider the sentence L6-1 in Table 8. In the listener’s TGHM, if the connection represented by the word “in” can be found between the nodes “Duke University” and “North Carolina,” the listener can respond with “Yes, it is” as feedback to the speaker. If such a connection cannot be found, the listener can reply with “No, it is not.”

- **Data Searching Sentence** (i.e., WH-Question Sentence):

When confronted with data searching sentences, the primary response is to furnish search results to the speaker. Take the sentence L6-2 in Table 8 as an example. In the listener’s TGHM, if a node exists at the other end of the edge indicated by the word “in,” it signifies a successful search mission, and the listener can provide the speaker with the lexical representation of that node. If a node is not found, the listener can reply with “I don’t know” or “I don’t have a clue” to inform the speaker that the search mission was failed.

### 4.3 Conceptual Interpretation

In this paper, we briefly introduce the new method mainly at the conceptual and practical levels. For the new concepts mentioned in the database-based method, we first give a rough interpretation under the previous theoretical frame for a better understanding. For example: 1) the TGHM, data chains, and the processing requests implicit in sentences are implicit knowledge; 2) the TGHM can also be explained as a semantic representation. 3) In the TGHM, the structure consists of computable edges can also be seen as the reasoning path, which giving our method the algebraic capacity to understand and generate a potentially infinite number of novel combinations from known components. 4) the reasoning process is also a practice of systematic generalization [8]. 5) Since, we have reclassified sentences according to their implicit processing requests, the previous classification of NLP tasks, e.g., dialogue and question answering, will be replaced. 6) The arrangement of the requisite parts in a sentence (see Table 6) and the sentence structures for encoding different data processing requests (see Table 7) are called syntax.

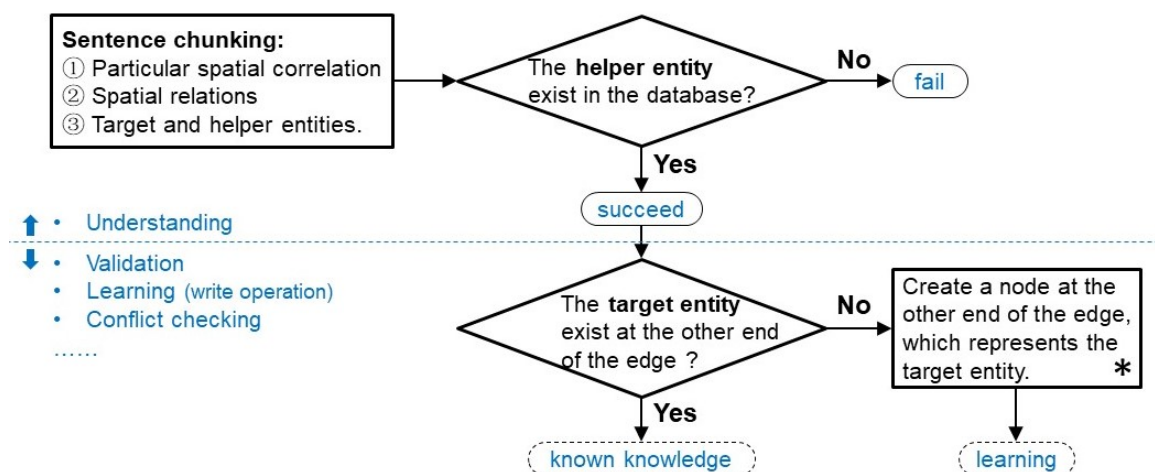


Figure 10: Processing flowchart of the data description sentences.

## 5 Logical Evaluation and Limitation

**Why do not provide the experimental evaluation results?** Approaches based on statistical learning methods usually consider their research objects to be random, stochastic, and indeterminate implicitly. Accordingly, performances on the benchmarks are used to evaluate the effectiveness of the approaches. However, when people say that things are random, stochastic, probabilistic, or due to change, what they really mean is that their outcomes are determined in part by a complex set of processes and deep structures that we are unable or unwilling to measure and will instead treat as random [6]. If we can find out these complex processes and deep structures of the research objects, their randomness is gone. Clearly, TGHM is a stable system, which describes the facts. Therefore, experimental test methods are not suitable for TGHM.

**Algebraic property** The edges in TGHM and the computability of these edges correspond to the characteristics of the algebraic structure: Group. This means that the edge relations between any two TGHM nodes are computable, which allows machines to generate sentences to describe any node in TGHM. Since TGHM (or databases) are our method's cornerstones, the database-based method's effectiveness is innate.

**Logical relations between Databases, NLG, NLU, and Learning tasks** We can summarize the logical processes in the database-based method as follows: The databases can be seen as a set of axioms; the NLG task is to derive propositions (i.e., sentences) from the axiom set; the NLU task is to verify sentences (propositions) with the axiom set, in which involves the validation and conflict checking processes. In an NLU process, if a given proposition is known to be true, the new information or knowledge brought by the proposition can be written into the database, further expanding the database (axiom set), which is a learning process. Translation tasks can also be performed using the database, the encoding table (see Figure 5), and corresponding encoding rules. The NLG, NLU, learning, and translation processes summarized above simulate how information and knowledge are processed in the human brain.

## 6 Conclusion

This work provided a new framework for solving NLP problems and discussed its potential in other AI problems (e.g., learning, translation). So, what exactly can we learn from the study of language? As we have learned in neuroscience, humans receive information through neural pathways such as eyes, ears, mouth, nose, etc., and then send this received information to the brain for hierarchical processing and storage. Although we cannot directly observe how this information is processed and stored in human brains, a small proportion of the information is encoded as natural language for external output. Thus, we can take natural language as a window to explore how information and knowledge are stored and processed in the human brain, which will lead to a brand new direction in AI research.

## References

- [1] Mark Bear, Barry Connors, and Michael A Paradiso. *Neuroscience: exploring the brain*. Jones & Bartlett Learning, 2020.
- [2] Emily M Bender and Alexander Koller. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198, 2020.
- [3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [4] N. Chomsky. *Syntactic Structures*. The Hague/Paris: Mouton, 1957.
- [5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [6] Nicholas J Gotelli, Aaron M Ellison, et al. *A primer of ecological statistics*, volume 1. Sinauer Associates Sunderland, 2004.
- [7] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.
- [8] Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pages 2873–2882. PMLR, 2018.
- [9] Ronald W Langacker. *Foundations of cognitive grammar: Volume I: Theoretical prerequisites*, volume 1. Stanford university press, 1987.
- [10] Ronald W Langacker. *Foundations of Cognitive Grammar, Vol. II: Descriptive Application*. Stanford University Press, 1991.
- [11] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*, 2020.
- [12] Baoyi Niu. Analysis of the methodological ideas of cognitive grammar. *Foreign Languages in China*, 2023.
- [13] Bernhard Schölkopf. Learning to see and act. *Nature*, 518(7540):486–487, 2015.
- [14] Leonard Talmy. *Toward a cognitive semantics, volume 1: Concept structuring systems*, volume 1. MIT press, 2003.
- [15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.